

# Tweet Characteristics Correlation to Engagement

Konner Sawyers

*College of Computer Science  
and Engineering*

*Marshall University*

Huntington, United States

Gabriel Smith

*College of Computer Science  
and Engineering*

*Marshall University*

Huntington, United States

Layne McNeely

*College of Computer Science  
and Engineering*

*Marshall University*

Huntington, United States

***Abstract*** – Social media is a useful set of tools for people and communities to promote their interest to a wider range of people. This includes one of the largest Twitter/X. Users of platforms like Twitter need to stay up to date on the most recent trends involving improving post engagement to ensure greater success in disseminating messages to a broad audience. We use a combination of API accessed web scrapers and Python Selenium scripts to retrieve and analyze many data points over the span of weeks. Our goal is to discover what factors in tweet creation and what times affect tweet success the greatest and to form suggestions for future users with the outcome.

## I. INTRODUCTION

Social media is one of the main sources of entertainment, news, entertainment, and advertising these days. There are many people who wish to use social networking tools for their personal or professional benefit and in light of that there is constantly a demand for knowing what does and does not work in the goal of increasing influence.

Once such network is X, we will refer to it by its previous name Twitter. Many people still use it even after its acquisition and therefore still require the knowledge of how to best use the system. Other studies [1] and [2] performed similar analysis but were aimed at a smaller target group.

Our goal is to provide a strategy to perform the best broadly on Twitter, meaning that the information provided can be used by anyone and not be targeted to any one community. This will be done by scraping tweets over the course of a couple weeks and measuring the success of each tweet to draw trends on what factors and time lead to greatest post success.

## II. METHODOLOGY AND RESULTS

### A. Equations and Definitions

Using Twitter itself as the source [3] engagement as a value must be defined in such a way that meets our requirements and fits our limitations.

$$= \frac{\text{Engagement Rate}}{\text{Impressions}} * 100\%$$

*(Likes + Replies + Quotes + Retweets)*

This is single tweet engagement rate where impressions are the number of views of a given tweet. It represents the likelihood of a tweet viewer performing an interaction with the tweet with a maximum value of 400% meaning that the viewer will perform four actions on that view or impression.

### B. Data Acquisition

Considering the recent changes around the acquisition of Twitter, now X, earlier this year it has become impossible to use Twitter's built in API to retrieve tweet data for less than \$100. So instead, an alternative had to be used and that came in the form of using the online platform Apify. In particular, the 'Tweet Flash – Twitter Scraper' Apify actor created by Jonathan Larson. With this we were able to collect thousands of posts without the hefty asking cost of Twitter's new API policy, however this is not without some drawbacks.

Although the actor could collect things such as likes, comments, images, and timestamps it was not capable of collecting account followers, likes, views, and location. This created a predicament in which we had all the engagement information but no data on impressions which prevented us from doing any calculations. This issue was resolved however with the creation of a Python powered Selenium script which allowed us to load Twitter's dynamic

webpages and manually scrape the information using the link provided in the initial data scrape. The now complete data was then used by the database to create a view using the provided *Engagement Rate* formula.

The queries used to get the post specified place, date, and time. The queries were generated using a script that iterated through days and the hours of each day with the location set to New York City with a 35-mile radius. The idea being that many tweets from different trends, organizations, individuals, and topics could be collected to generate broad trends with no bias towards any poster or community. There is also the upside of knowing where all the post came from, allowing all times to be adjusted towards EST (Eastern Standard Time).

### C. Q1 Analysis

Are there factors of tweet creation, such as hashtag count, video inclusion or mentions, that can impact the likelihood of better engagement rates?

This problem was approached from multiple angles by measuring factors in an isolated setting and still integrated with other factors. To keep things clear the list of factors measured were:

- Video
- External Link
- Images
- Hashtags
- Mentions
- Length

Effects were calculated by presence and by count where appropriate. Also, only tweets that had more than one hundred views or impressions were used. Anything less and the engagement rate could be heavily skewed and ruin the results.

The way the success was gauged was by classifying the engagement rate percentile of each tweet. These percentiles were based on the averages of the data we collected and were set to:

Below 25 <sup>th</sup> Percentile	Poor/Very Low
Between 25 <sup>th</sup> and 50 <sup>th</sup> Percentile	Low
Between 50 <sup>th</sup> and 75 <sup>th</sup> Percentile	Average
Above 75 <sup>th</sup> Percentile	High

Figure 1 Percentile ranges category key

Our first test involved taking tweets that did have a certain feature and comparing the performance classification to tweets that did not have that given feature and putting the results in tables separated by the feature analyzed. This resulted in data showing how certain features can hurt or improve performance overall.

TWEET PERFORMANCE BASED ON FEATURE INCLUSION

Video							
	Poor	Low	Average	High	Average %	High %	Average+
Yes	219	260	247	191	26.94	20.83	47.77
No	1901	1856	1881	1919	24.89	25.39	50.28
External Link							
Yes	288	112	109	119	17.36	18.95	36.31
No	1832	2004	2019	1991	25.73	25.38	51.11
Hashtag							
Yes	185	188	246	316	26.31	33.8	60.11
No	1935	1928	1882	1794	24.96	23.8	48.76
Images							
Yes	403	420	513	614	26.31	31.49	57.8
No	1717	1696	1615	1496	24.75	22.93	47.68
Mentions							
Yes	201	191	218	238	25.71	28.07	53.78
No	1919	1925	1910	1872	25.05	24.55	49.6

Figure 2 Tweet performance based on the inclusion/exclusion of a given feature.

According to the results in figure 3, the inclusion of mentions, images and hashtags improve the performance of post without those features with hashtags having the greatest change in performance at a 10-point increase in the high percentage category over its counterpart. On the other hand, inclusion of a video and or external link negatively affects the engagement rates with video inclusion dropping high engagement rate by 4.56-points from 25.39 and external links dropping the high engagement rate by 6.43 from 25.38. This implies that posters are better off avoiding the usage of those two pieces of content.

The second set of tests is in two parts, the data was classified based on feature combinations. This was done quantitatively, where the count of each feature was used, and then another done as a binary classification where the classification was defined by the presence of that feature or not. Then only the classifications that had more than one hundred results were kept ensuring a proper sample size. All the data

here was generated in table and clustered bar graph pairs to give a better visualization of the trends.

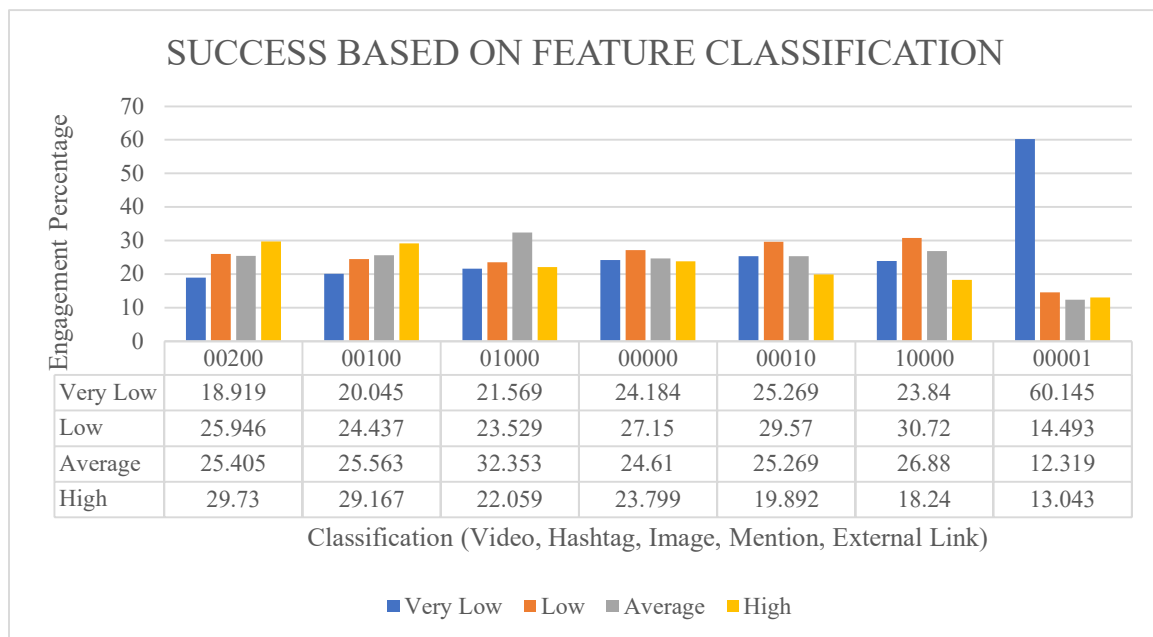


Figure 3 Classification based on the count of each feature.

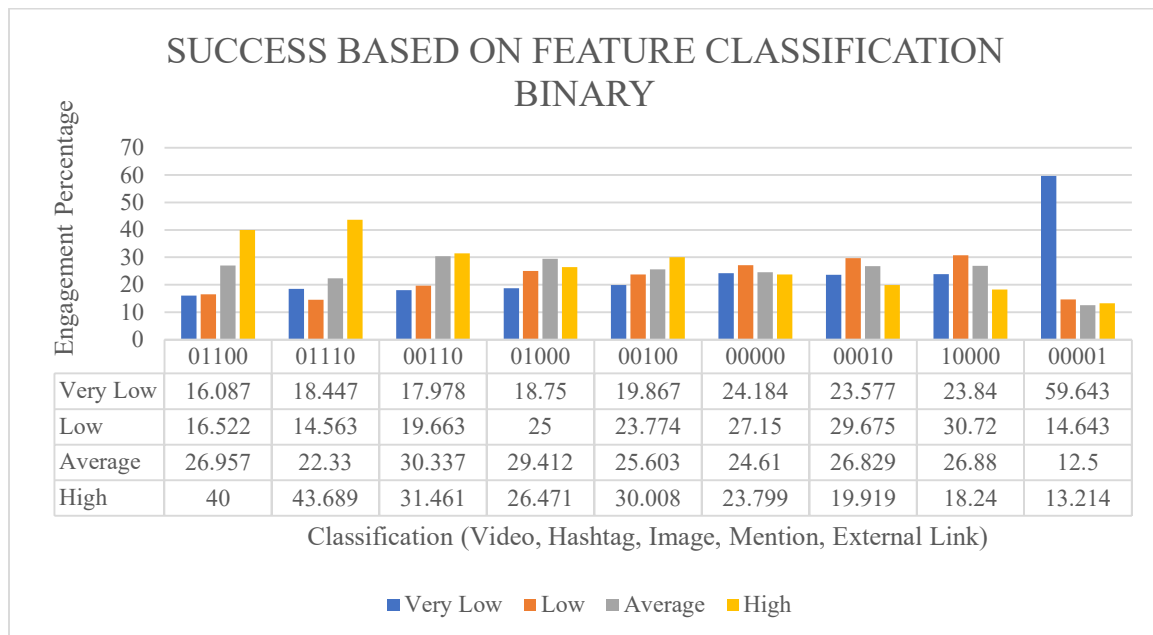


Figure 4 Classification based on if the feature exists in a post or not. Maximum of  $5^2$  categories possible.

Figure 4 and figure 5 further expand upon the information from figure 3. The results in figure 3 show that hashtags provide the largest benefit overall to high engagement rate however when sorted out and observed on their own has poorer performance compared to the other combinations of features. It

seems that hashtags provide the increased engagement rate previously shown only when in combination with other features such as mentions or pictures.

Figure 4 on its own doesn't reveal too much other than posting an external link or video on its own will

basically guarantee poor performance with them presenting with a combined low + very low engagement rate of 74.286% and 53.72% respectively meaning at least those should be avoided.

Figure 5, however reveals more details as to what makes a good combination. Overall, it seems that the two best options are having at least one hashtag, image, and mentions for a high rate of 43.689% or excluding the mention for a rate of 40%. These values exceeded the average high rate of figure 4 (27.42%) by 16.269% and 12.58% respectively.

The final analysis for tweet features is regarding the tweet length. Tweets were only used if equal to or under the character limit of 280. The reason being that post can only be made on paid accounts which isn't specifically what we are interested in. The range of 0 to 280 characters was broken up into a histogram of sorts with the x-axis values being selected by taking the square root of 280 and rounding up to the nearest factor, that being 20 characters. Then tweets in the range have their engagement percentages categorized before being graphed.

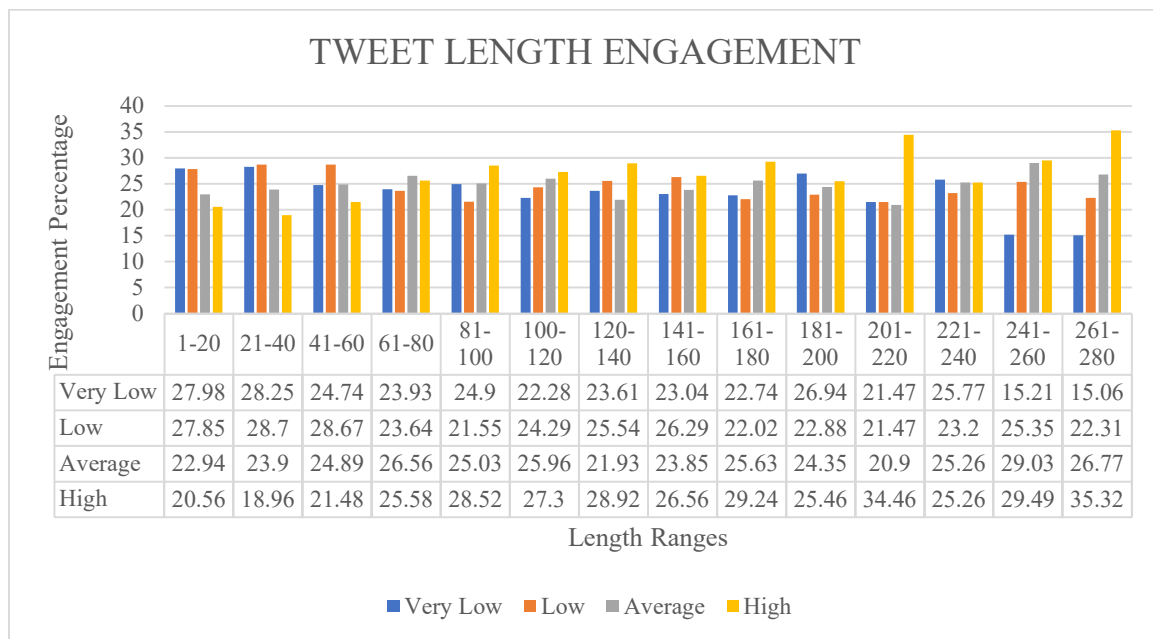


Figure 5 Engagement rate regarding the character length/count.

The results for tweet length seem to be a little inconsistent. The highest values are in the 201-200 at 34.46% and 261 - 280 at 35.32%. The worst performance tweets were everything 1 – 60. Everything in that range deviated more than 5-points below the mean whereas beyond that even if a range was below the mean, such as 221 – 240 it was a less than 1.7-point deviation.

#### D. Q2 Analysis

Does the time of day or the day of the week for a post affect the engagement performance?

The process for the analysis here is a bit simpler compared to the previous set. First off, we repeated the process of only retrieving post where the views are greater than one hundred. The good part of how the data was initially retrieved is that there is guaranteed data for every hour of every day scraped ensuring a good sample size for each. Post retrieved are sorted yet again according to the key provided in fig 2. Posts are then independently classified based on day posted and hour posted, and the data is displayed in table and clustered bar graph pairs.

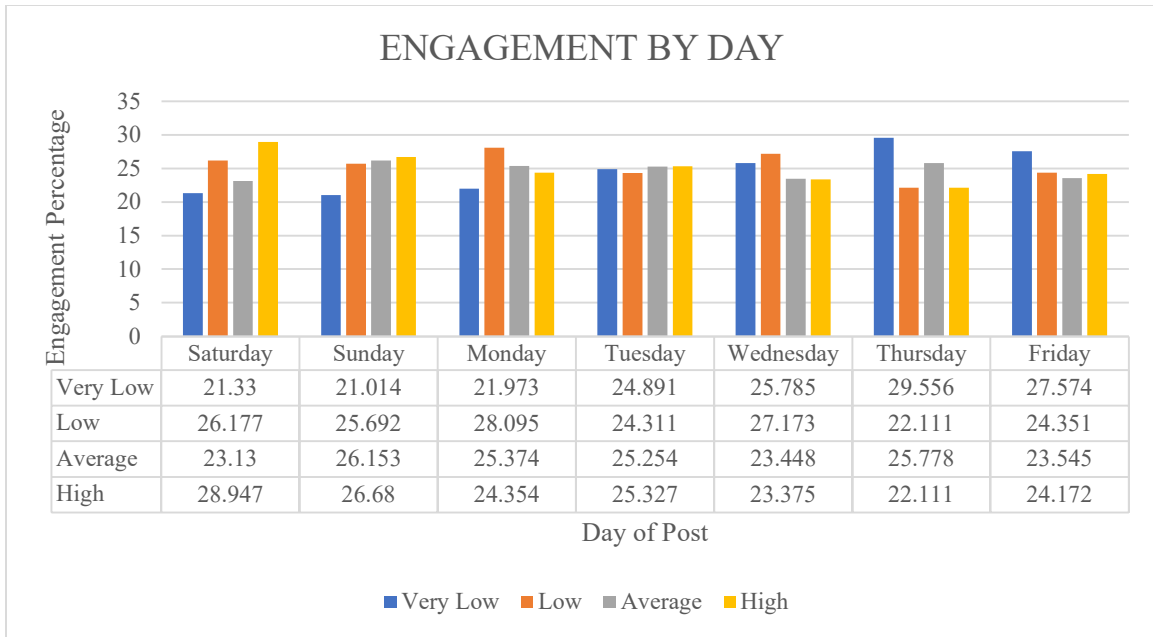


Figure 6 Tweet engagement rate sorted by the day posted.

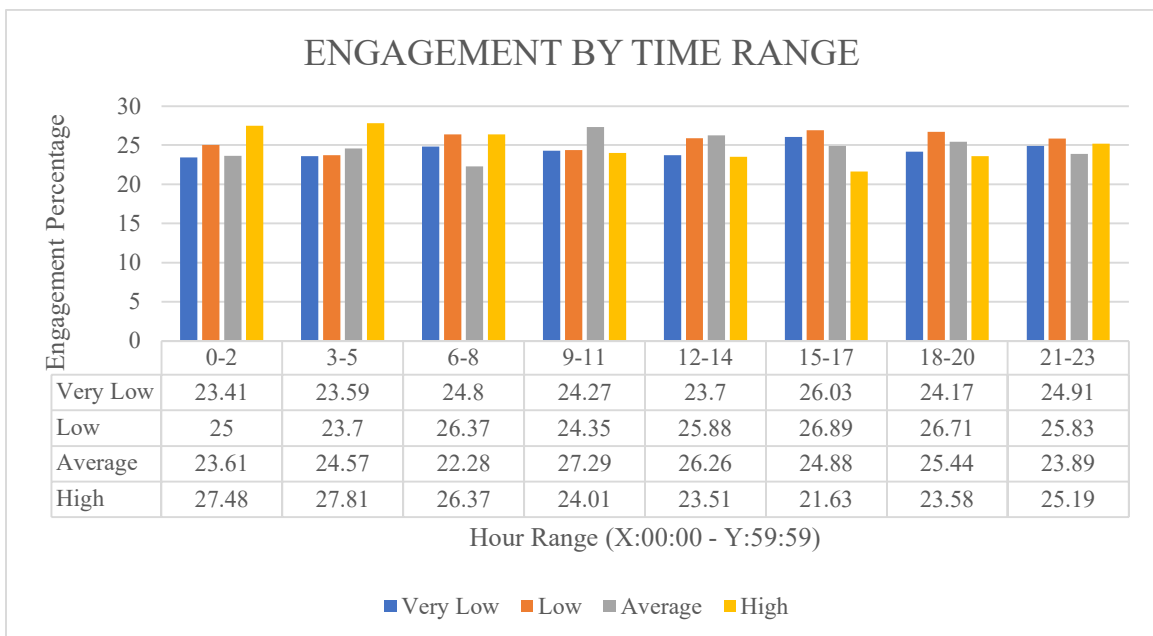


Figure 7 Tweet engagement rate based on the hour range posted.

The results for question 2 are a little less obvious compared to the trends analyzed in the previous question due to the smaller levels of deviation, however they still exist. In figure 7 the mean high engagement rate rests at 24.995% and the only days that surpass that are Saturday, Sunday, and Tuesday in descending order with Tuesday not even being .5-points greater than the mean.

For figure 8 the mean high engagement is 24.9475% meaning that the time range of 9:00 – 20:59 falls below that value. For the time ranges that did exceed the value of the mean both 0:00 – 2:59 and 3:00 – 5:59 were greater than the mean by more than 2-points.

### III. RESULTS DISCUSSION & RELATED WORK

The analysis of tweets based on its features excluding the length seems conclusive. It falls in with results discovered by others such as the case with Wadhwa's group [1] with the inclusion of hashtags and images. Ours just expanded the recommendation to mentions as well.

Analysis on tweet length seems to be inconsistent, the reason for this could be the sample size needed to be larger or that length is more dependent on community or trend. Overall though it seems that tweets should be longer than 60 characters to avoid the drop off seen with tweets below that length.

Regarding the time and day of posts, the weekends did better overall than the weekdays, with Saturday being the absolute best, and the best time range to post was around midnight to 6pm. This falls close to the findings of others [4] where they found the best time to be 2 hours around midnight with a decrease in activity past 4pm.

Overall considering how results seem to somewhat overlap the findings of others helps to validate what we found. However, some improvements to the process can be made. For example, the initial gathering process can be replaced with a self-made script for gathering posts to prevent the occurrence of incomplete fields. This allows for the removal of the extra data retrieval step increasing accuracy and scraping speed. Another improvement to the overall process is to use more tweets if done again to further increase the number of categories for analyzing such as with figure 4 and figure 5. Due to the lack of many diverse tweet structures figure 4 is missing many

combinations which could grant greater insight if present.

### IV. CONCLUSION

To improve tweet performance, include features such as mentions, image(s), or hashtags preferably in combination with each other for the best outcomes. To avoid wasting efforts its best to post in the early morning, around midnight to 6 pm, with the best days to post being the weekend.

## References

- [1] V. Wadhwa, E. Latimer, K. Chatterjee, J. McCarty, and R. T. Fitzgerald, "Maximizing the tweet engagement rate in Academia: Analysis of the AJNR Twitter Feed," *American Journal of Neuroradiology*, vol. 38, no. 10, pp. 1866–1868, Jun. 2017, doi: 10.3174/ajnr.a5283.
- [2] X. Han, X. Gu, and S. Peng, "Analysis of Tweet Form's effect on users' engagement on Twitter," *Cogent Business & Management*, vol. 6, no. 1, Jan. 2019, doi: 10.1080/23311975.2018.1564168.
- [3] "Engagement API: Interpreting the metrics" accessed: October 2023.  
<https://developer.twitter.com/en/docs/twitter-api/enterprise/engagement-api/guides/interpreting-metrics>
- [4] B. Shahzad, E. Alwagait, "Best and the Worst Times to Tweet: An Experimental Study," in *Proc. of the 8th International Conference on MANAGEMENT*,